Projeto Final

Disciplina Mineração de Dados em Biologia Molecular 2012

O objetivo deste projeto é utilizar os conceitos de Mineração de Dados vistos durante a disciplina de forma prática e combinada, utilizando, para isso, ou um conjunto de dados da área de biologia que ele tenha acesso, de pesquisa feita por ele na USP, ou 3 (três) conjuntos de dados do UCI Machine Learning Repository. O trabalho pode ser feito de forma individual apenas.

Para o trabalho, o aluno deve preparar o conjunto de dados de sua pesquisa no na USP ou os 3 (três) conjuntos de dados disponíveis publicamente relacionados à biologia do site da UCI, dentre os conjuntos (da área de aplicações Life Sciences):

- Ecoli
- p53 Mutants
- Parkinsons
- Primary Tumor
- PubChem Bioassay Data (um dos 21 conjuntos de dados)
- Soybean (Small)
- Yeast

Deve escolher também 3 classificadores implementados na ferramenta WEKA para a realização dos experimentos (escolher dentre: SVMs, redes neurais MLP, J48, naive bayes, redes neurais RBF, JRip, SinmpleCart, logistic

regression e KNN). Escolhidos o conjunto de dados e os classificadores, as seguintes tarefas devem ser realizadas com ele:

- Para cada conjunto de dados selecionado, descrever as principais características em uma tabela (número de exemplos, número de atributos numéricos, número de atributos categóricos, número de exemplos por classe).
- Converter cada conjunto para o formato ARFF (utilizado pela ferramenta WEKA).
- Selecionar, no WEKA, manualmente os atributos que considerar relevante para o problema dentre os atributos do conjunto original (pode acontecer de todos os atributos serem selecionados).
- Comparar 2 técnicas de amostragem do WEKA com o uso do conjunto completo para todos os algoritmos de classificação utilizados no projeto (3, se individual e 6, se em dupla).
- Comparar desempenho do conjunto de dados original com 2 técnicas de seleção baseadas em filtro e 2 baseadas em *wrapper* para os algoritmos de classificação utilizados. Usar opção com validação cruzada.
- Usar *scatter plot* e *boxplot* para ilustrar a distribuição dos valores nos atributos selecionados por uma das técnicas anteriores. Para esses mesmos dados, apresentar a média e o desvio padrão das medidas de média, moda, mediana, curtose e obliqüidade observadas para cada um dos atributos selecionados.
- Se os dados forem desbalanceados, comparar 3 técnicas de balanceamento. Se não forem, realizar um experimento reduzindo aleatoriamente pela metade o número de exemplos de uma das classes e realizar, nesse novo conjunto, a comparação de 3 técnicas de balanceamento.

- Usar teste de hipóteses para comparar o desempenho dos classificadores entre eles. Fazer isso:
    - com e sem atributos selecionados (por uma das técnicas apenas).
    - Com e sem as classes balançadas (por uma das técnicas apenas).
- Escrever um relatório, descrevendo a base de dados e os algoritmos de aprendizado de máquina utilizados, relatando o que foi feito e mostrando tabelas, gráficos e análise dos resultados.

Entrega: 30/11/2012